Reconstruction of global monthly upper-level temperature and geopotential height fields back to 1880

Griesser, T ; Brönnimann, S ; Grant, A ; Ewen, T ; Stickler, A ; Comeaux, J

Abstract: This work presents statistically reconstructed global monthly mean fields of temperature and geopotential height (GPH) up to 100 hPa for the period 1880–1957. For the statistical model several thousand predictors were used, comprising a large amount of historical upper-air data as well as data from the earth's surface. In the calibration period (1958–2001), the statistical models were fit using the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) as the predictand. After the weighting of the predictors, principal component (PC) analyses were performed on both the predictand and predictor dataset. Multiple linear regression models relate each principal component time series from the predictand with an optimal subset of principal component time series from the predictor. To assess the quality of the reconstructions, statistical split-sample validation (SSV) experiments were performed within the calibration period. Furthermore, the reconstructions were compared with independent historical upper-air and total ozone data. Based on the SSV experiment, this study obtained good reconstructions for temperature and GPH in the Northern Hemisphere; however, the skill in the tropics and the Southern Hemisphere was much lower. The reconstruction skill shows a clear annual cycle with the highest values in January. The lower levels were better reconstructed except in the tropics where the highest levels showed the best skill. With the inclusion of a considerable amount of upper-air data after 1939 the skill increased substantially. The fields were analyzed for selected months in the 1920s and 1930s to demonstrate the usefulness of the reconstructions. It is shown that the reconstructions are able to capture regional-to-global dynamical features.

DOI: https://doi.org/10.1175/2010JCLI3056.1

Reconstruction of Global Monthly Upper-Level Temperature and Geopotential Height Fields Back to 1880

THOMAS GRIESSER, STEFAN BRÖNNIMANN, ANDREA GRANT, TRACY EWEN, AND ALEXANDER STICKLER

Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

JOEY COMEAUX

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 27 January 2009, in final form 15 March 2010)

ABSTRACT

This work presents statistically reconstructed global monthly mean fields of temperature and geopotential height (GPH) up to 100 hPa for the period 1880–1957. For the statistical model several thousand predictors were used, comprising a large amount of historical upper-air data as well as data from the earth's surface. In the calibration period (1958–2001), the statistical models were fit using the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) as the predictand. After the weighting of the predictors, principal component (PC) analyses were performed on both the predictand and predictor dataset. Multiple linear regression models relate each principal component time series from the predictand with an optimal subset of principal component time series from the predictor. To assess the quality of the reconstructions, statistical split-sample validation (SSV) experiments were performed within the calibration period. Furthermore, the reconstructions were compared with independent historical upper-air and total ozone data. Based on the SSV experiment, this study obtained good reconstructions for temperature and GPH in the Northern Hemisphere; however, the skill in the tropics and the Southern Hemisphere was much lower. The reconstruction skill shows a clear annual cycle with the highest values in January. The lower levels were better reconstructed except in the tropics where the highest levels showed the best skill. With the inclusion of a considerable amount of upper-air data after 1939 the skill increased substantially. The fields were analyzed for selected months in the 1920s and 1930s to demonstrate the usefulness of the reconstructions. It is shown that the reconstructions are able to capture regional-to-global dynamical features.

1. Introduction

For the study of interannual-to-decadal climate variability in the twentieth and the late nineteenth centuries, a variety of global gridded datasets of different variables on a monthly or daily basis at the surface are available [e.g., the second Hadley Centre's monthly historical mean sea level pressure data (HadSLP2; Allan and Ansell 2006); European and North Atlantic mean sea level pressure (EMSLP; Ansell et al. 2006); Climatic Research Unit Temperature and Salinity, version 2.1 (CRU TS 2.1; Mitchell and Jones 2005); the Hadley Centre Climatic Research Unit, version 3, surface temperature (HadCRUT3; Brohan et al. 2006); the Hadley Centre Sea Ice and SST (HadISST; Rayner et al. 2003); and the Extended Reconstruction of SST (ERSST; Smith and Reynolds 2004)]. Gridded datasets going further back in time (e.g., Luterbacher et al. 2002, 2004; Casty et al. 2005; Pauling et al. 2005; Xoplaki et al. 2005) are on a more regional scale (e.g., Europe) and have less temporal resolution.

Although many phenomena can be addressed to some extent based on surface data, an interpretation of these phenomena requires information on atmospheric circulation, which necessarily involves features at upper levels. Datasets consisting of direct upper-air measurements, such as radiosondes or pilot balloons, exist for the second half of the twentieth century [e.g., the Comprehensive Aerological Reference Data Set (CARDS;

Corresponding author address: Thomas Griesser, Institute for Atmospheric and Climate Science, ETH Zurich, Universitätstrasse 16, 8092 Zurich, Switzerland.
E-mail: thomas.griesser@alumni.ethz.ch

DOI: 10.1175/2010JCLI3056.1

Eskridge et al. 1995; see also Lanzante et al. 2003); the Hadley Centre's radiosonde temperature product (HadRT; Parker et al. 1997); and the Integrated Global Radiosonde Archive (IGRA; Durre et al. 2006)]. Parts of these datasets, supplemented with additional information from the surface and satellites, were assimilated into weather prediction models to generate probably the most widely used 3D datasets: the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) and the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalyses (Uppala et al. 2005; Kistler et al. 2001). Although the continuously updated NCEP–NCAR reanalysis now provides data for the past 60 yr (1948–2008), there is still a general lack of knowledge about the variability of the upper-level circulation in earlier times and on interdecadal time scales.

Several authors have tried to fill in this gap [for an early attempt, see Kington (1975)]. Klein and Dai (1998) presented a method to statistically reconstruct 700-hPa geopotential heights (GPH) for North America, extending to the Pacific and Atlantic, using surface air temperature (SAT) and sea level pressure (SLP) data. Schmutz et al. (2001) reconstructed 700-, 500-, and 300-hPa GPH for the European and eastern North Atlantic region based on SLP, SAT, and precipitation (RR) data back to 1900. Gong et al. (2006) derived 500-hPa GPH for the Northern Hemisphere based on SLP and SAT fields back to 1871. All the studies mentioned earlier have one major shortcoming: they do not include any upper-air measurements. Brönnimann and Luterbacher (2004) pointed to the availability of upper-air measurements before 1948, and they reconstructed 700-, 500-, 300-, and 100-hPa GPH and temperature back to 1939 using surface and newly available upper-air measurements. In this paper, we extend and refine this approach, earlier described by Cook et al. (1994) and Jones et al. (1987) as orthogonal spatial regression, and present statistical reconstructions of GPH and temperature for three regions: the Northern Hemisphere (NH; 15°–90°N), the tropics (TP; 20°S–20°N), and the Southern Hemisphere (SH; 15°–90°S). The dataset consists of monthly reconstructions on the levels 850, 700, 500, 300, 200, and 100 hPa for the period 1880–1957 to allow for a seamless connection to the ERA-40 reanalysis. Note that we propose to use the regions separately where possible. Global fields can be produced by using a linear combination in the overlapping latitudes, but we leave it to the user to choose the appropriate weights as this will strongly depend on the application.

Another method of filling this gap is data assimilation. A new reanalysis (G. P. Compo et al. 2010, unpublished manuscript) has been produced based only on sea level pressure data and monthly sea surface temperature and sea ice fields (hence, no upper-air information) reaching back to 1908. The two datasets are complementary and allow interesting comparisons (although they are not independent). Detailed comparisons of these two and other datasets will be presented elsewhere.

This paper is organized as follows: in section 2, the data used for reconstruction and validation are presented; section 3 describes the reconstruction method; validation results are shown in section 4; some analyses of reconstructed fields, demonstrating the potential value of these data, are presented in section 5; and conclusions are presented in section 6.

2. Data

a. Terminology

For the reconstruction we define two major time periods: the calibration/validation period (1957–2002) and the reconstruction period (1880–1957). The statistical model relates a predictand (Y) to a predictor (X) dataset. In this reconstruction approach, the predictand consisted of the upper-level temperature and GPH fields. The predictor comprised upper-level and surface-based measurements. A third dataset consisting of independent upper-air data was required for validation in the reconstruction period; this is in addition to an initial quality assessment, which was already performed in the calibration/validation period based on a split-sample validation (SSV) approach (see section 3c). In the following sections, the datasets are briefly described and their quality discussed. For a further discussion of each dataset, the reader is referred to the cited literature.

b. Predictand data in the calibration/validation period

As a predictand, a long, global, and homogeneous 3D dataset was required. The two datasets used in the majority of cases are the ERA-40 and NCEP–NCAR reanalyses. NCEP–NCAR starts in 1948 and is continuously updated to the present (Kistler et al. 2001). ERA-40 starts in 1957 and ends in 2002 (Uppala et al. 2005). The operational forecasting system and assimilation procedure in NCEP–NCAR was designed in the mid-1990s, whereas the core of ERA-40 was developed after 2000. Therefore, the two reanalyses belong to two different generations. In direct comparisons, ERA-40 clearly outperforms NCEP–NCAR (Simmons et al. 2004; Santer et al. 2004; Bengtsson et al. 2004). For this reason we choose ERA-40 as predictand for the reconstruction.

Although most deficits apparent in NCEP–NCAR were removed in ERA-40, some problems remain unsolved

and some new problems were added. In the Southern Hemisphere, the data coverage is still poor in the early years, especially before 1967 (Uppala et al. 2005). Small jumps in the mean temperatures in the troposphere are present, resulting from differences in the bias correction of satellite measurements, with the largest inhomogeneity expected around 1975–76. In the presatellite years, the extratropical Southern Hemisphere exhibits a cold tropospheric bias (Bengtsson et al. 2004). In the same years, a cold bias in winter and springtime in the Antarctic lower stratosphere is apparent. Hence, there is an inhomogeneity within the predictand data that is also evident at the surface (Simmons et al. 2004).

The disadvantage of the shorter time period, compared to NCEP–NCAR, was expected to be at least compensated by the increased data quality. Furthermore, despite the problems with ERA-40 mentioned previously, there were other good reasons to use it as predictand. Our reconstruction approach primarily focused on spatial variability patterns for the whole troposphere and the lowermost stratosphere and was therefore less affected by inhomogeneities in either a subregion or a specific layer. Additionally, the month-to-month variability is large relative to the observed jumps. Inhomogeneities in the predictand dataset only affect the quality of the reconstruction to the extent to which they project onto patterns of variability that occur naturally. Also, they do not introduce trends in the reconstruction period. The quality of the reconstruction can be assessed with a statistical bootstrap procedure in the calibration period and additionally with the independent validation data in the reconstruction period.

In our case, we used monthly mean fields of GPH and temperature at the 850-, 700-, 500-, 300-, 200-, and 100-hPa levels (termed Z850, T850, Z700, T700, etc.) interpolated to an equal-area grid. Hence, the number of grid points on a latitudinal circle decreases toward the poles. The distance on a longitudinal circle was kept constant with a resolution of 2.5°. We have a maximum number (144) of grid points at the equator, equivalent to a resolution of 2.5°. This number decreases toward the poles according to the cosine of the latitude.

c. Predictor data

The predictor data were divided into two major groups: surface data and upper-air data. The surface data again consisted of gridded SLP (HadSLP2; Allan and Ansell 2006) and the Goddard Institute for Space Studies homogenized surface station temperature data (GISSTEMP; Hansen et al. 1999). The SLP dataset was incorporated ''as is'' with a spatial resolution of 5° by 5° and spanning from 1880 to 2002.

For the surface temperature predictors, stations with high data quality and good spatial and temporal coverage are preferred. Therefore, the GISSTEMP station network was reduced according to the following criteria. First, all stations with less than 90% of possible data available in the historical period were eliminated. Second, we calculated the Pearson correlation between the temperature anomalies of each individual station and ERA-40 925-hPa temperature anomalies interpolated to the station location. Stations with a correlation <0.8 were removed. Third, because the United States still showed an overrepresentation relative to other regions, which is potentially problematic with regard to the weighting, the station network over the United States was further reduced. U.S. stations with an incomplete record in the twentieth century were discarded. Based on the criteria just described, a global subset of 760 total stations was extracted covering the period from 1880 to 2002. The location of the surface temperature stations and the temporal evolution of the number of predictors are presented in Fig. 1 [note that similar findings result when using gridded temperature data such as the Climatic Research Unit Temperature dataset, version 3 (CRUTEM3v) instead of station data]. After subtracting the annual cycle based on the period 1961–90, the data were standardized, and the few remaining missing data points in the calibration period in the extracted surface station network were filled with similarly standardized 925-hPa anomalies from ERA-40 to have a complete data series. Brönnimann and Luterbacher (2004) showed that this is justified by the high median correlation of 0.85 between the reanalysis and the station series. The standardization is necessary to control the information that goes into the principal component (PC) analysis and to get rid of the units.

For the upper-air data we distinguished between measurements taken by radiosondes, kites, aircraft, and pilot balloons. All upper-air measurements were from the period before 1958 and originated from many different sources. The radiosonde data were collected from digital archives and were compiled and quality controlled (Grant et al. 2009b). They were supplemented with additional historical radiosonde, aircraft, and kite measurements which were processed at the Swiss Federal Institute of Technology (ETH) Zurich (Brönnimann 2003a,b; Ewen et al. 2008b). In addition, reevaluated upper-level wind data from the global TD52 and TD53 pilot balloon datasets provided by NCAR (available online at http://dss.ucar.edu/docs/papers-scanned/papers.html, datasets RJ0167 and RJ0168) and from the African pilot balloon dataset of Météo-France were used.

The radiosonde data were quality checked and homogenized (see Grant et al. 2009b for a detailed overview).

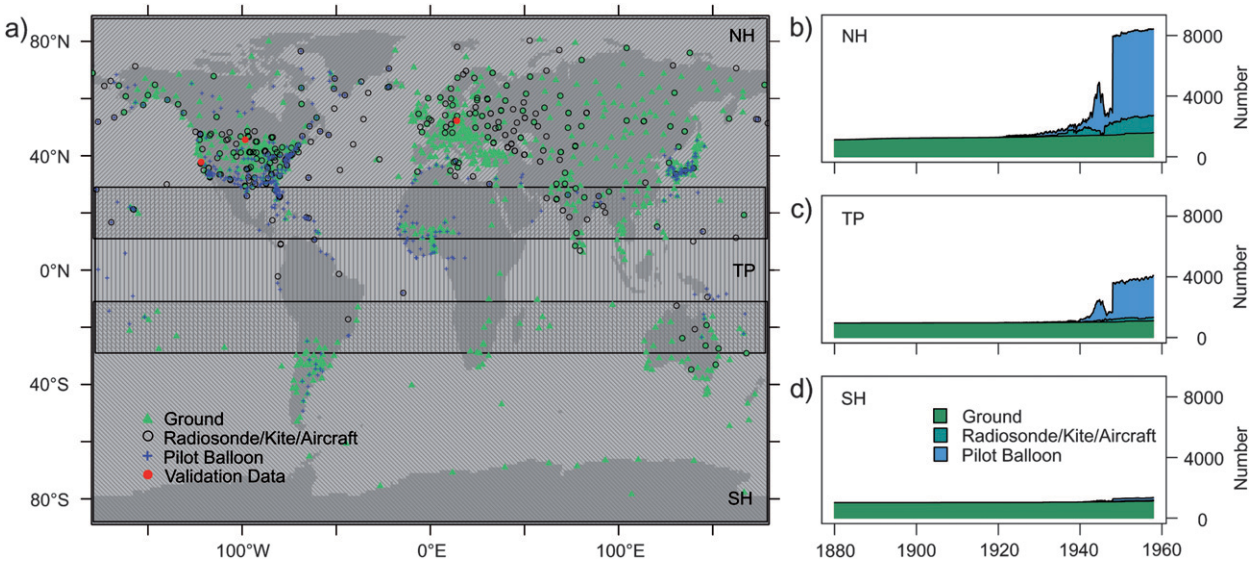

FIG. 1. (a) Map of surface and upper-air stations used as predictors. The three regions (NH: 10°–90°N; TP: 30°S–30°N; SH: 10°–90°S) used for the reconstruction can be seen. Green triangles represent surface temperature stations, blue crosses denote pilot balloon stations, black circles denote upper-air series taken by radiosondes, kites, or aircrafts, and red circles are upper-air stations used for the validation. (b),(c),(d) Number of available predictors from 1880 to 1957 separated by measurement platforms for the regions NH, TP, and SH, respectively.

Pilot balloon data were also quality assessed (Stickler et al. 2010). In the cases where it was not clear if the station should be accepted or rejected, generally because of a too weakly correlated reference series, the variance and the mean of the historical period were plotted against the same variables from ERA-40 at the same location to assess the quality in a climatological way. Station series with a bias of more than 2 standard deviations, or a difference in the variance of more than 1.5 standard deviations between the historical period and the reanalysis, were rejected if the historical time series was longer than 1 yr. Whenever the majority of the levels from a station showed inconsistency with the reanalysis, the complete station was removed.

The location of all upper-air predictors retained for the reconstructions as well as the measurement platforms is shown in Fig. 1. Globally, 15 394 upper-air predictor variables in the historical period were used for the reconstruction (7752 kite/aircraft or radiosonde and 7642 pilot balloon).

All upper-level series cover only a part of the historical period and most do not reach the present time or have long gaps, often because stations were relocated or closed or the measurement platform changed—for instance, no kite data are available after the 1930s. Therefore, ERA-40 was used to supplement all historical upper-level series after 1958. Note that most of the data series after 1958 were assimilated into ERA-40. Hence, where observations would be available, they are supposed to be well represented by ERA-40, and where they are not available, ERA-40 is a good estimate for most regions. The only exceptions were the TD52 and TD53 datasets after 1948, which were rigorously quality checked in a previous study (Ewen et al. 2008a) using the NCEP–NCAR reanalysis. Data from that study were used after supplementation with NCEP–NCAR after 1948.

The interpolated reanalysis has less variability than observational data as it does not represent local features and smoothes out random errors and to some extent biases. To account for this, we perturbed the interpolated reanalysis data by a bias and a noise component.

The ''bias'' is defined per station and variable and is constant over time and across levels (the GPH bias increases with altitude as described below). For each station and variable, the amount of the bias is sampled from a normal distribution with standard deviations of 0.5°C for temperature, 0.7 m s^{-1} for wind, and 7.5 (at 850 hPa) to 20 gpm (at 100 hPa) for GPH. Note that the bias can have any sign. The noise is completely random, that is, it is not constant over time or across levels nor does it depend on the bias. The noise also is sampled from a normal distribution with standard deviations of 1.1°C for temperature, 1.1 m s^{-1} for wind, and 11.5–53 gpm for GPH. The standard deviations of the bias and noise components were estimated based on the quality assessment of the data (see Brönnimann 2003a; Grant et al. 2009b). After perturbation, all predictor variables

were standardized and expressed as anomalies with respect to the 1961–90 annual cycle.

The data availability for any given month in the historical period was much more limited than suggested by Fig. 1. Except for the SLP data, all data series had longer or shorter gaps in the historical period. The earliest upper-air series used for the reconstruction started in 1920 and a large amount of data was confined to the lower troposphere. The coverage was much better for the continents and the Northern Hemisphere. For the tropics and the Southern Hemisphere, the coverage was poor and upper-air data were available only from the beginning of the mid-1930s.

d. Validation data

For the purpose of validation, some upper-air stations were retained and not used for the reconstruction. Stations were selected according to the following criteria: first, the stations had to cover as much of the historical period as possible, preferably with no gaps; and second, to keep the validation of the reconstruction independent from the quality control procedure of the predictors, only stations that did not need any correction were used. Based on these criteria, three stations were withheld: Oakland (United States), Ellendale (United States), and Lindenberg (Germany; see Fig. 1 for their exact positions). Lindenberg was the upper-air station with the longest available record, going back to 1905. Furthermore, the reconstruction was compared with independent reconstructions from Brönnimann and Luterbacher (2004) and Schmutz et al. (2001).

In addition to historical upper-air data, we used historical total ozone data to assess the reconstruction (see Brönnimann and Staehelin 2004 for a brief description of the technique of validating upper-level reconstructions with total ozone). At midlatitudes, total ozone is known to be well correlated with meteorological variables in the tropopause region. Historical total ozone data were available for time periods (e.g., the 1920s) and regions (e.g., Australia or China) for which no upper-air data were available and thus allowed some conclusions about the skill of the reconstructions in these cases. We used monthly mean values of total ozone at Arosa (Switzerland), 1926–2002 (Staehelin et al. 1998); Tromsø (Norway), 1935–72 (Hansen and Svenøe 2005); Oxford (United Kingdom), 1924–75 (Vogler et al. 2007); New York (United States), 1941–44, and Shanghai (China), 1932–42 (Brönnimann et al. 2003); and Canberra (Australia), 1929–32 (unpublished, reevaluated as in Brönnimann et al. 2003). Where no station data were available after 1978, Total Ozone Mapping Spectrometer (TOMS) total ozone data (version 8) were used to supplement the series.

TABLE 1. Average radius (km) for a given level (L0: surface; L1: 250–3000 m or 925–700 hPa; L2: 3001–6000 m or 699–500 hPa; L3: 6001–9000 m or 499–300 hPa; L4: above 9000 m or below 300 hPa) beyond which the spatial correlation drops below 0.5, which defines the influence radius of the stations.

Level/variable	Temperature	GPH	Wind
L4	1529	1483	1311
L3	1379	1425	1267
L2	1398	1448	1142
L1	1421	1487	1017
L0	1266	—	—

3. Reconstruction method

a. Weighting scheme

The available historical predictors were unequally distributed in space. In general, there was an overrepresentation of the earth's surface compared to the middle and upper troposphere and continents were better covered than oceans. This fact potentially results in an unintentional focus on small-scale variability near the surface and over landmasses. Hence, the station series must be weighted to better represent the whole variability present in the predictor dataset.

As a first step, all data series were assigned to an altitude level (L0: surface; L1: 250–3000 m or 925–700 hPa; L2: 3001–6000 m or 699–500 hPa; L3: 6001–9000 m or 499–300 hPa; L4: above 9000 m or below 300 hPa). For the reconstruction of the Northern Hemisphere (15°–90°N) predictors from north of 10°N were included, for the tropics (20°S–20°N) predictors from 30°S to 30°N were used, and for the Southern Hemisphere (15°–90°S) all predictors south of 10°S were used. Second, within each level and for the variables GPH, temperature, and wind (u and v winds were treated as a single variable), the average 0.5 decorrelation distance was calculated, yielding an estimation of the ''influence radius.'' Influence radii for each variable and level are given in Table 1. The weight for each individual station and variable was the inverse of the number of all available stations with information from the same variable within the influence radius. Finally, the weights were adjusted such that the overall weight of a variable in a level was proportional to the total area covered by all the influence radii combined. The covered area as a percentage of the total area for each region and variable is shown in Fig. 2, and a map showing the coverage for the month indicated in Fig. 2 is given in Fig. 3. This procedure was repeated for each time step. Within the surface level (L0), 50% of the weight was attributed to SLP and 50% to the surface station temperature field. The SLP field was additionally area weighted.

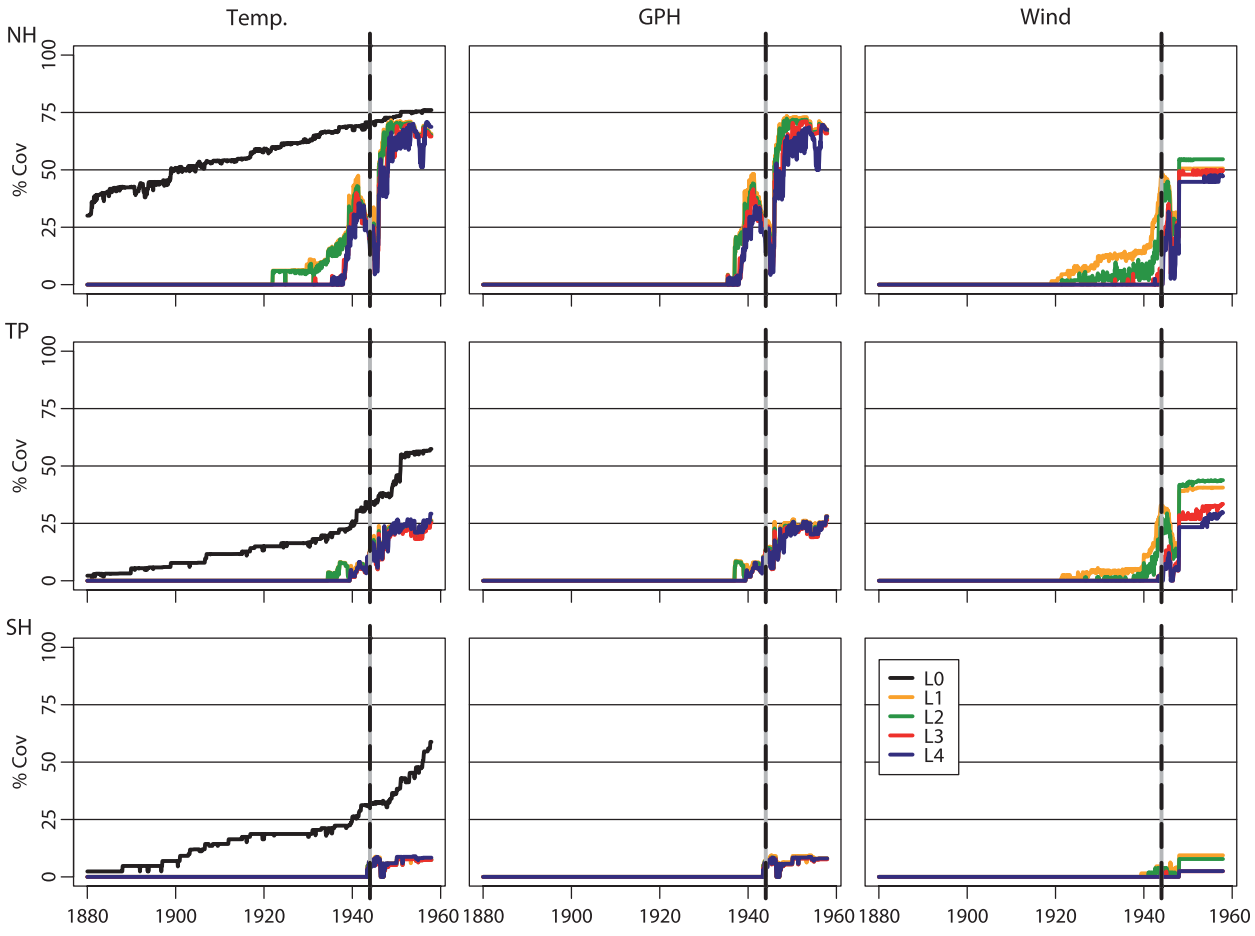

FIG. 2. Temporal evolution of the percentage of total area that was covered for a given variable [(left) temperature, (middle) GPH, and (right) wind], region [(top) NH, (middle) TP, and (bottom) SH]. Levels (L0–L4) are indicated by color. The dashed vertical line indicates January 1944, the month for which the total area coverage is presented in Fig. 3.

b. Statistical model: Setup

After the regridding of the predictand to an equal-area grid and after the perturbation of the predictor, the regression model was set up. Here, the model is defined as a set of regression equations linking the predictors and predictands for a specific month. The statistical model was fit in the calibration period and the derived relation was applied to the reconstruction period. The approach used here was based on a PC regression model, similar to Brönnimann and Luterbacher (2004), which is also known as orthogonal spatial regression (Cook et al. 1994; Jones et al. 1987).

As described in the data section, the predictor network in the historical period changes over time and longer or shorter gaps existed in some predictors. To make use of all available data, a separate statistical model for each historical month was created.

To calibrate a model for a specific month in the past, a 3-month moving window centered on the associated calendar month was used. For the reconstruction of January 1944, for example, all data from the months December, January, and February in the calibration period were selected. In a further step, only those predictor series that were available in the defined month in the reconstruction period were selected for the calibration period. The extracted subset of predictor variables was multiplied by the square root of the weighting field pertaining to the specified historical month (for the weighting field, see section 3a). Next, one PC analysis was performed on the predictand dataset (standardized, all variables and levels combined) and a second on the predictor subset. Each predictand PC time series was then expressed as a linear combination of an optimal subset of predictor PC time series using linear regression (least squares estimator). To obtain the best model, only a certain number of predictor and predictand PCs were retained. The retained variance was varied between 70% and 98% (independently on both the predictor and the predictand side), and the subset with the

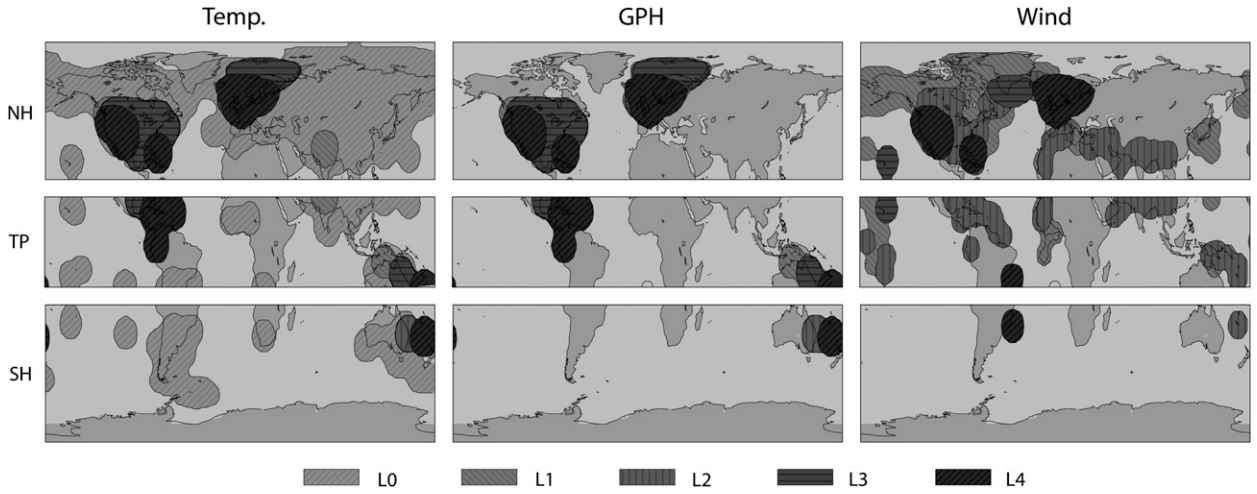

FIG. 3. Map showing the covered areas, defined by the influence radii, for January 1944 and (left to right) the variables GPH, temperature, and wind. The different shadings indicate the levels L0–L4 (see text).

best performance was chosen for the reconstruction (where best performance is measured according to the split-sample validation; see section 3c). The predictand PCs were generated by applying the coefficients to the corresponding predictor PCs in the reconstruction period. The reconstructed anomaly field is a function of the reconstructed predictand PCs that were retained and the predictand PC scores from the calibration period. Finally, the standardization procedure was inverted and the data were regridded to a 2.5° by 2.5° grid.

To reconstruct monthly values for 78 yr (1880–1957), 936 individual models were formed for each of the three regions. Every model consisted of an iteratively solved set of regression equations based on least squares.

c. Statistical model: Validation

Before the final reconstructions were carried out, different sensitivity experiments were performed concerning the robustness of the reconstruction and weighting method, the definition of appropriate regions for the reconstruction (NH, TP, and SH), and whether to reconstruct all levels and variables together or separately (not shown). To determine the potential benefit from the inclusion of upper-air predictors, reconstructions with only surface data were also conducted.

The reconstructions were validated by using the SSV technique, a special case of a cross validation. The calibration period for the final reconstruction (1958–2001) was split into a calibration part and a validation part for the SSV model. The statistical model was derived from the data in the SSV calibration period and tested in the independent SSV validation period. This procedure was repeated twice with different time periods. The model was fit either in the period 1958–87 or 1972–2001 and tested in the period 1988–2001 or 1958–71, respectively. The potential skill of the model was measured with the reduction of error statistic (RE; Cook et al. 1994) defined as

$$\mathrm{RE} = 1 - \frac{\sum_t (x_{\mathrm{rec}} - x_{\mathrm{obs}})^2}{\sum_t (x_{\mathrm{null}} - x_{\mathrm{obs}})^2}, \qquad (1)$$

where t is time, x_{rec} is the reconstructed value, x_{obs} is the observed value, and x_{null} is the null hypothesis. For reconstructed anomalies, the null hypothesis corresponds to a 0 anomaly from the long-term mean annual cycle (1961–90). Values of RE can be between $-\infty$ and 1 (perfect reconstruction). An RE of 0 is indicative of a reconstruction not better than climatology, whereas an $\mathrm{RE} > 0$ points to a model with predictive skill. Because of the stochastic properties, RE values can be above zero by chance. Therefore, we consider reconstructions useful if RE values are above 0.2. This approximately corresponds to R^2 equal to 0.2–0.25 (see Brönnimann and Luterbacher 2004). Because the validation period in the SSV procedure is 14 yr long, Eq. (1) sums over 14 time steps. The result of each SSV experiment is a spatial field of RE values on the predictand grid. For the model validation it is useful to aggregate the information into a single number. As the RE skill score has a fixed upper boundary at 1, distributions of RE values tend to be skewed. In this case, the appropriate location estimator is the RE median. For the selection of an optimal subset of predictand and predictor PCs, the RE median over the entire field was calculated and maximized. For the analysis of the fields, usually the average RE value from the two split-sample validations was given.

In the work by Ewen et al. (2008a), a subsample of our predictor dataset was used for the reconstruction of an upper-level index (the Pacific–North American pattern) using a very similar approach. In that study, additional validations were performed in a surrogate climate using NCAR Community Climate System Model, version 3 (CCSM3) output (Collins et al. 2006). The reconstructions in the surrogate climate showed almost identical skill to that obtained from the SSV in the reanalysis data. The conclusion was that this reconstruction method is robust in the model environment. It would be beneficial to repeat the same validation experiments for the full reconstruction presented here, but that is beyond the scope of this paper.

4. Validation results

a. Sensitivity studies

In Table 2, the results of three sensitivity studies are presented. The reconstruction procedure is repeated without weighting, without adding noise in the calibration period, and without weighting or noise. For comparison the results of the final reconstructions are shown as well. The performance of the different models is measured based on the SSVs using the RE and coefficient of efficiency (CE) skill scores. CE has a similar definition to RE, but it takes into account a potential bias between calibration and validation period and is therefore slightly more pessimistic (see Cook et al. 1994). Results are given for distinct fractions of explained variance on the predictor and predictand side and for the two split-sample validations individually or conjoined. Because neither the addition of noise nor the weighting of predictors changes the outcome dramatically, the statistical model can be considered as robust. The highest RE and CE values are only slightly affected by adding noise or weighting. At the same time, the lowest skill score values increase significantly when using weighting and perturbation, most visible in the RE and CE mean of the temperature reconstruction. Therefore, we have to assume that both weighting and perturbation improve the statistical model in the way that relevant modes in the predictors are better identified (better separation between noise and signal), while at the same time an overfitting of the model is prevented.

Beside the later-used principal component analysis (PCA) method, the reconstruction procedure was repeated with canonical correlation analysis (CCA; see Cook et al. 1994) for selected months. With respect to the considered digits in the RE and CE values, the results looked similar and therefore are not shown in the table.

b. Split-sample validation

Results from the SSV are summarized in the form of time series of the field median value of RE for a region (NH, TP, and SH) or as maps showing the RE field for a specific level and month or period. As the SSV used only 30 yr for calibration (rather than the full data), the results might provide a rather pessimistic estimation of the skill. On the other hand, the SSVs themselves are used in the optimizing procedure and therefore might be too optimistic.

The SSV showed distinctly different results for the three independently reconstructed regions. Based on the SSV (see Fig. 4 for the median of the RE fields for selected levels), the performance of the statistical model was best for the Northern Hemisphere, intermediate for the tropics, and poor for the Southern Hemisphere. Nevertheless, some findings concerning the quality of the reconstructions are valid in general. The RE_{median} time series from all regions and levels show an annual cycle, most evident in the GPH fields in the Northern Hemisphere, with the highest RE_{median} usually found in January in all regions. Geopotential height was generally better predicted than temperature, and lower levels were better reproduced than higher levels. The inclusion of upper-air predictors increased the reconstruction skill in all regions and on all levels, with the most pronounced for higher levels and temperature.

The RE_{median} values in the Northern Hemisphere showed a clear annual cycle with a maximum in January and a minimum in June. This phenomenon is known and has been described previously (Cook et al. 1994; Jones et al. 1987; Brönnimann and Luterbacher 2004). In the period without (before 1920) or with few (before 1939) upper-air data, the skill of the reconstruction was good during the Northern Hemisphere winter for GPH at all levels and for temperature in the lower levels (T850, T700, and T500). Furthermore, the SSV results suggest poor predictability for the T300, T200, and T100 levels before 1939.

The inclusion of a considerable number of upper-air predictors after 1939 increased the reconstruction skill substantially. The effect was largest for temperature and at the higher levels, bringing the RE_{median} of the different levels closer together. With RE_{median} values between 0.6 and 0.8 during the Northern Hemisphere winter, good reconstructions were found for GPH after 1940, whereas reasonable to good reconstructions are obtained for GPH during summer and temperature the whole year round. The RE_{median} in the tropics showed a clear annual cycle with a superimposed semiannual signal. Normally, RE_{median} was highest in January. For GPH at the higher levels, this peak was shifted to February. A secondary

TABLE 2. Detailed results of the validation experiment showing RE and CE values for January 1940 for the Northern Hemisphere. Displayed are the mean and the median for the 700-hPa level for both temperature and GPH. The values are calculated for each split-sample validation individually (SSV1 or SSV2) or conjoined (SSV1 + SSV2). The experiment was carried out for certain fractions of retained variance in the predictors (% explained variance of X) and predictands (% explained variance of Y). The validation procedure was repeated using different setups of the statistical model. (a) The SSV results shown are based on the statistical model that was used for the final reconstructions, therefore, with perturbed and weighted predictors like outlined in the paper. The experiment is repeated (b) without adding noise on the predictors in the calibration period, (c) without the weighting of the predictors, and (d) without noise and weighting. For (b)–(d), only RE and CE values for the best model (90% Expl. Var in X and Y) are shown. Note that the best model selection is based on the RE_{median} for the conjoined six levels and therefore is not identical with the highest RE_{median} shown in (a).

		GPH												Temperature											
		RE						CE						RE						CE					
(a)		Median			Mean			Median			Mean			Median			Mean			Median			Mean		
% Expl. Var. X	% Expl. Var. Y	SSV1 + SSV2	SSV1	SSV2	SSV1 + SSV2	SSV1	SSV2	SSV1 + SSV2	SSV1	SSV2	SSV1 + SSV2	SSV1	SSV2	SSV1 + SSV2	SSV1	SSV2	SSV1 + SSV2	SSV1	SSV2	SSV1 + SSV2	SSV1	SSV2	SSV1 + SSV2	SSV1	SSV2
70	70	0.55	0.56	0.55	0.48	0.48	0.48	0.49	0.43	0.53	0.39	0.32	0.46	0.39	0.42	0.37	0.31	0.37	0.25	0.32	0.32	0.31	0.24	0.28	0.20
	80	0.57	0.57	0.57	0.51	0.52	0.49	0.51	0.48	0.56	0.42	0.37	0.47	0.40	0.44	0.34	0.32	0.40	0.24	0.33	0.36	0.28	0.25	0.31	0.19
	90	0.58	0.58	0.58	0.53	0.53	0.52	0.52	0.48	0.57	0.44	0.37	0.50	0.42	0.45	0.40	0.34	0.40	0.29	0.35	0.35	0.35	0.27	0.30	0.25
	95	0.55	0.58	0.52	0.50	0.52	0.49	0.50	0.49	0.52	0.41	0.36	0.46	0.41	0.45	0.36	0.34	0.40	0.28	0.34	0.36	0.31	0.27	0.31	0.23
	98	0.54	0.57	0.47	0.48	0.52	0.43	0.47	0.47	0.47	0.39	0.37	0.41	0.39	0.44	0.34	0.33	0.40	0.26	0.33	0.36	0.30	0.25	0.29	0.21
80	70	0.60	0.59	0.61	0.52	0.50	0.54	0.54	0.47	0.60	0.43	0.35	0.52	0.42	0.44	0.41	0.36	0.41	0.31	0.36	0.36	0.36	0.29	0.32	0.26
	80	0.63	0.62	0.64	0.56	0.54	0.57	0.57	0.53	0.63	0.47	0.39	0.55	0.44	0.48	0.39	0.37	0.44	0.31	0.38	0.40	0.34	0.30	0.35	0.25
	90	0.64	0.62	0.66	0.58	0.55	0.61	0.59	0.52	0.65	0.49	0.39	0.59	0.47	0.48	0.45	0.39	0.44	0.34	0.40	0.41	0.40	0.32	0.35	0.30
	95	0.64	0.64	0.64	0.57	0.56	0.59	0.60	0.57	0.63	0.49	0.41	0.56	0.45	0.48	0.40	0.37	0.44	0.30	0.39	0.42	0.36	0.30	0.36	0.25
	98	0.62	0.63	0.61	0.54	0.57	0.52	0.57	0.55	0.59	0.46	0.43	0.50	0.44	0.49	0.37	0.35	0.45	0.24	0.38	0.42	0.34	0.28	0.36	0.19
90	70	0.60	0.60	0.61	0.52	0.52	0.53	0.54	0.48	0.60	0.43	0.36	0.50	0.42	0.43	0.41	0.34	0.40	0.28	0.36	0.36	0.36	0.27	0.31	0.23
	80	0.66	0.65	0.66	0.59	0.58	0.59	0.61	0.55	0.65	0.50	0.44	0.57	0.46	0.48	0.44	0.39	0.45	0.32	0.41	0.41	0.40	0.32	0.36	0.27
	90	0.70	0.68	0.71	0.64	0.61	0.67	0.66	0.59	0.70	0.57	0.48	0.65	0.51	0.50	0.52	0.44	0.47	0.40	0.47	0.45	0.48	0.38	0.39	0.36
	95	0.70	0.71	0.70	0.63	0.61	0.64	0.66	0.62	0.69	0.55	0.48	0.62	0.51	0.52	0.49	0.42	0.47	0.36	0.46	0.46	0.46	0.35	0.39	0.31
	98	0.68	0.70	0.65	0.60	0.61	0.59	0.63	0.62	0.64	0.53	0.49	0.57	0.47	0.51	0.43	0.36	0.46	0.26	0.42	0.44	0.40	0.30	0.38	0.21
95	70	0.60	0.59	0.61	0.52	0.51	0.52	0.55	0.47	0.60	0.43	0.35	0.50	0.42	0.43	0.40	0.33	0.40	0.27	0.35	0.36	0.35	0.26	0.31	0.21
	80	0.65	0.65	0.66	0.57	0.57	0.57	0.60	0.55	0.65	0.49	0.42	0.55	0.45	0.47	0.43	0.38	0.45	0.31	0.40	0.41	0.39	0.31	0.36	0.26
	90	0.70	0.70	0.71	0.63	0.61	0.65	0.66	0.61	0.70	0.55	0.47	0.63	0.51	0.51	0.51	0.42	0.47	0.37	0.46	0.45	0.47	0.36	0.39	0.32
	95	0.71	0.71	0.70	0.62	0.61	0.62	0.66	0.64	0.68	0.54	0.48	0.60	0.49	0.50	0.47	0.39	0.45	0.33	0.44	0.44	0.45	0.32	0.37	0.27
	98	0.70	0.71	0.69	0.59	0.58	0.60	0.66	0.63	0.67	0.52	0.46	0.59	0.47	0.49	0.44	0.32	0.42	0.23	0.41	0.42	0.39	0.25	0.34	0.17
98	70	0.60	0.59	0.61	0.52	0.51	0.52	0.55	0.47	0.60	0.42	0.35	0.50	0.41	0.42	0.40	0.33	0.40	0.27	0.35	0.35	0.35	0.26	0.31	0.22
	80	0.66	0.65	0.67	0.57	0.57	0.57	0.61	0.55	0.66	0.48	0.41	0.55	0.45	0.47	0.43	0.38	0.44	0.31	0.40	0.41	0.39	0.31	0.36	0.26
	90	0.70	0.69	0.71	0.63	0.61	0.64	0.67	0.61	0.70	0.55	0.46	0.63	0.50	0.50	0.51	0.41	0.46	0.36	0.46	0.45	0.47	0.35	0.39	0.31
	95	0.70	0.70	0.70	0.61	0.61	0.62	0.66	0.62	0.68	0.53	0.46	0.60	0.49	0.50	0.48	0.38	0.45	0.31	0.45	0.45	0.45	0.31	0.37	0.26
	98	0.70	0.72	0.68	0.59	0.59	0.59	0.66	0.64	0.67	0.52	0.47	0.57	0.48	0.51	0.43	0.32	0.43	0.20	0.43	0.45	0.39	0.25	0.36	0.15
(b)																									
90	90	0.71	0.70	0.72	0.63	0.61	0.65	0.67	0.62	0.71	0.56	0.49	0.64	0.49	0.50	0.49	0.42	0.47	0.37	0.44	0.43	0.45	0.36	0.39	0.32
(c)																									
90	90	0.73	0.73	0.73	0.65	0.64	0.67	0.70	0.67	0.72	0.59	0.52	0.65	0.50	0.50	0.50	0.40	0.44	0.36	0.44	0.43	0.44	0.33	0.36	0.31
(d)																									
90	90	0.72	0.72	0.72	0.63	0.62	0.64	0.68	0.65	0.71	0.56	0.50	0.63	0.45	0.46	0.45	0.34	0.41	0.28	0.39	0.38	0.40	0.27	0.32	0.22

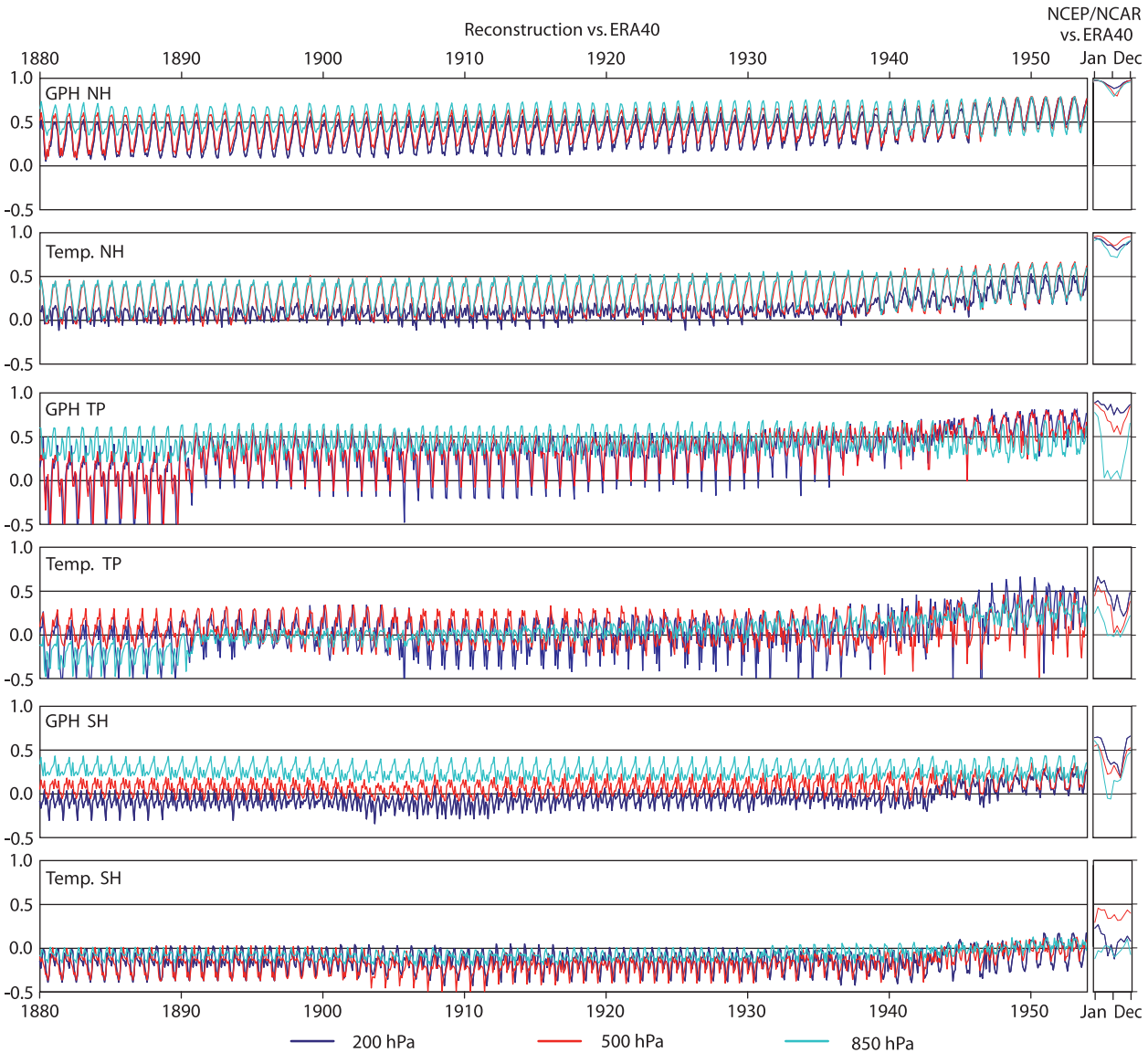

FIG. 4. (left) Time series of RE_{median} as a function of (top to bottom) variables and regions with height level shown in color from the split-sample validation. (right) RE values calculated from comparing the ERA-40 and NCEP–NCAR reanalysis for the period 1957–71 (summed over the 14 yr with respect to the calendar month).

maximum was reached in July. The annual cycle of the RE_{median} for the tropics is larger than for the Northern Hemisphere.

Whereas the reconstructions for Z850 and Z700 were reasonable to good from 1880 through the 1950s, RE_{median} remained low for GPH in the higher levels and for all temperature levels before the mid-1930s. The inclusion of upper-air data after 1935 raised the skill of the reconstruction model, especially for GPH during summer. The RE_{median} increased most for the highest levels for both GPH and temperature. In the Z100 level (not shown), RE_{median} values of 0.85 were reached in the 1950s. Although the annual cycle of the RE_{median} was reduced by introducing upper-air predictors, very low skill can be observed even in the 1950s in April–May or August–September for some years. In the Southern Hemisphere, the skill was very low before the 1950s. The only exceptions were the Z850 and Z700 levels, which showed an acceptable RE_{median} of 0.2–0.4. Although the few upper-air predictors raised the skill considerably, the RE_{median} remained on a low level even in the 1950s. Surprisingly, the maximum of the RE_{median} was reached in January.

To get an estimate of the maximum expected skill, the RE_{median} was calculated for a comparison of the two

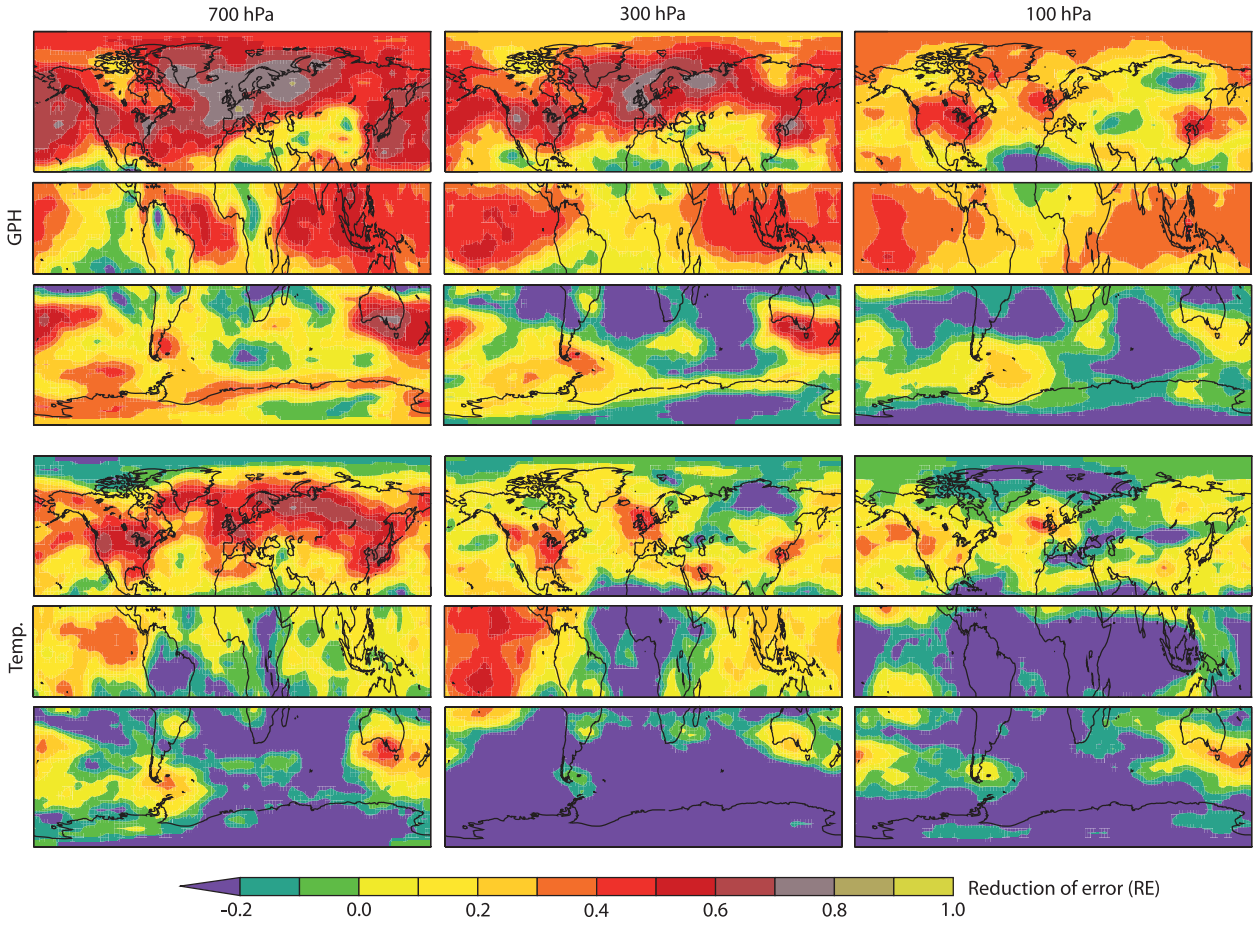

FIG. 5. Maps showing fields of RE_{median} for the period January 1900–December 1904 for the three regions for (top three rows) GPH and (bottom three rows) temperature on the levels (right to left) 700, 300, and 100 hPa.

reanalyses in the period 1957–71. For the sake of comparability with the results obtained from the reconstructions, the reanalyses' anomalies were interpolated to the equal-area grid described for the predictand. For the calculation of RE [see Eq. (1)], x_{obs} was defined as ERA-40 and x_{rec} as NCEP–NCAR. The results are shown in Fig. 4 (right). In the Northern Hemisphere, both reanalyses show excellent agreement for GPH and temperature at all levels. Although in the tropics the skill for the higher levels and GPH was still good, it dropped considerably for temperature on all levels and for GPH on the lower levels, especially from May to August, probably because of small-scale processes not resolved in the reanalyses, like convection. In the Southern Hemisphere, skill was low except for GPH from October to March and for 500-hPa temperature.

While the RE_{median} is a good measure for the overall fit of the model, the spatial details deserve more attention. Even if the field median RE value was low, some regions still showed good skill, depending on the season and available station network. Figures 5 and 6 show fields of RE_{median} values calculated over the periods 1900–04 and 1940–44, respectively. The earlier period represents a time when no upper-air predictors were available, whereas in the later period the station network was relatively well developed. As already seen from the analysis of the RE_{median} time series, RE values for GPH were clearly better than for temperature and lower levels were better reproduced than higher levels (except for the tropics). This fact also appeared in the RE patterns. For the Northern Hemisphere, independent of whether upper-air data were available or not, RE is high over North America, Europe, and East Asia–Japan. Poor reconstructions were found over parts of central and northeast Asia, North Africa, parts of the Atlantic and Pacific subtropics, and the north polar region. Upper-air data increased the skill in general and especially over central Asia and northwestern North America. In the tropics, skill was high over the Pacific except for the T100 level. Lower skill was found over

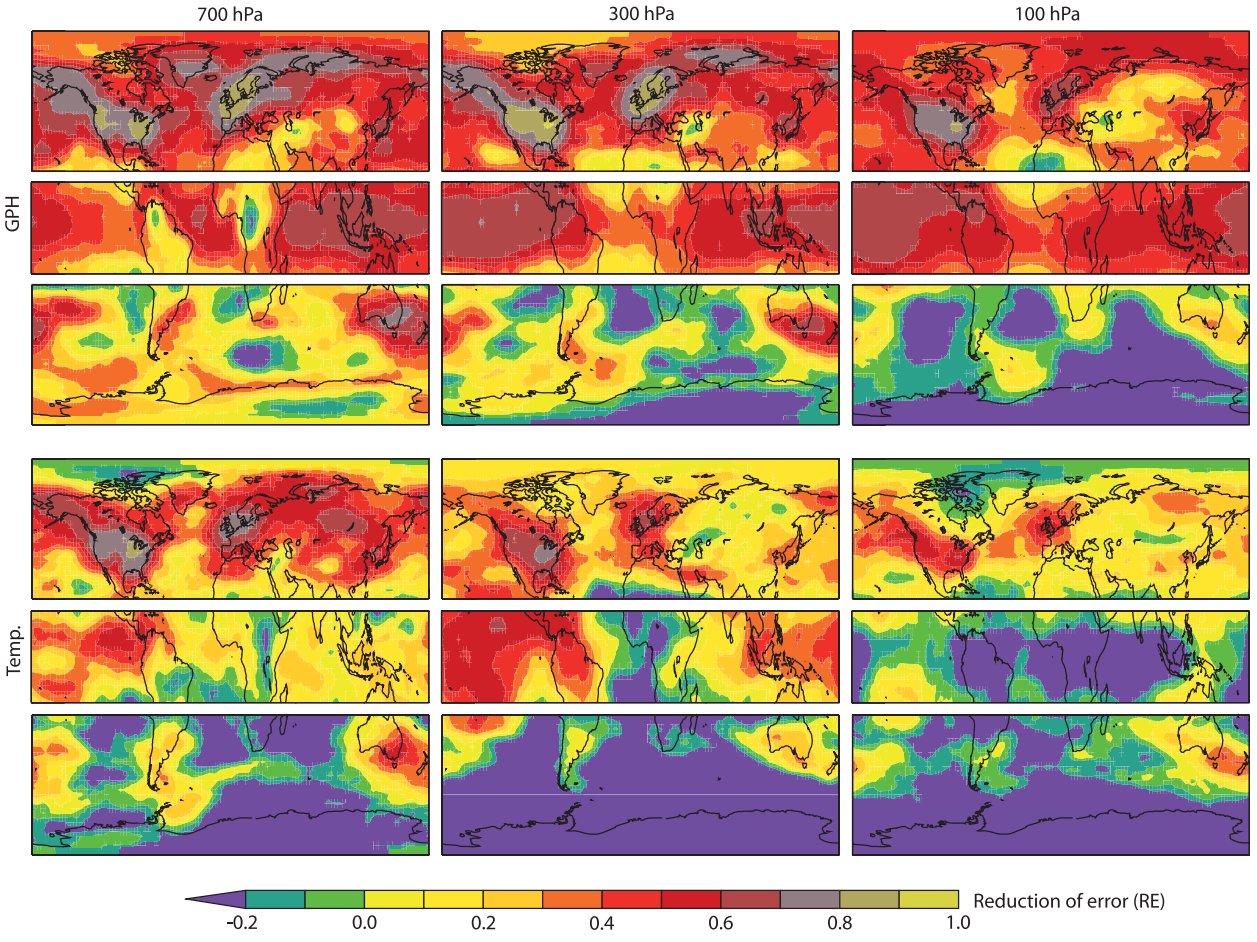

FIG. 6. As in Fig. 5, but for the period January 1940–December 1944.

the Atlantic, central Africa, and Central America for some levels.

The T100 level showed exceptionally low RE values, particularly spanning from the Indian Ocean over Africa to Central America. Inclusion of upper-air data generally increased the RE. For the Southern Hemisphere, although RE_{median} time series implicated low skill, good skill was found over the southeastern part of Australia.

In Figs. 7 and 8, the Z300 RE fields for January and July in 1904 and 1944, respectively, are shown. While in 1904 the skill remained high year-round for temperature and GPH over central North America, the northern Atlantic, western Europe, the tropical Pacific, the Indian Ocean, and parts of Australia, some regions showed much lower RE values in July. Most striking are the low values over the northwestern Pacific, parts of Asia, central Africa extending to the eastern Atlantic, and the whole Antarctic. Because of the inclusion of upper-air data, the skill in 1944 was generally better year-round. Some regions showed an increase in the RE values above the average, especially in July. For example, the SSV exhibited skill for northwestern North America and the central and northern parts of Australia in 1944, whereas there was no skill at all in 1904.

To separate the influence of including upper-air predictors from the effect of a changing surface station network, sensitivity experiments with only surface predictors were performed. Figure 9 shows RE fields for Z700, T700, Z300, and T300 for the Northern Hemisphere with (upper panels) and without (lower panels) upper-air data. While the inclusion of upper-air data increased the RE values for GPH on a global scale, the effect on temperature was more regional and is best visible over northern America and northern Europe.

c. *Validation with historical upper-air data and independent reconstructions*

In addition to the SSV, the reconstructions were compared with independent historical upper-air data (i.e., data not used for the reconstruction), as this is the only way to validate the final reconstruction. RE values were calculated in the same way as the SSV. For comparison,

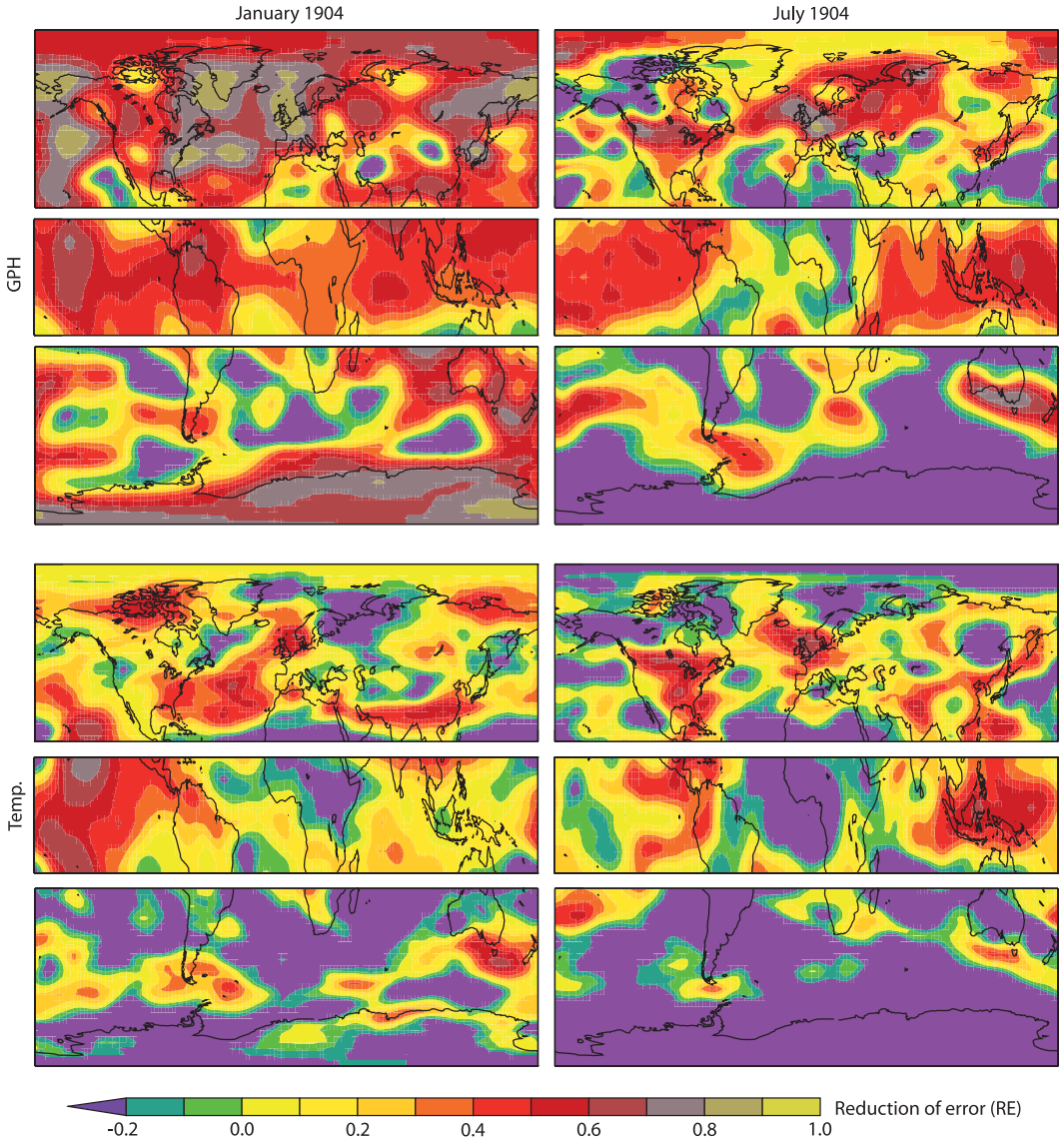

FIG. 7. Maps showing RE fields for (top six panels) GPH and (bottom six panels) temperature on the 300-hPa level for (left) January and (right) July 1904.

the data were pooled for each level. RE values were between 0.57 and 0.81 for GPH and between 0.33 and 0.55 for temperature. The corresponding correlations were between 0.75 and 0.81 for GPH and between 0.48 and 0.75 for temperature. Surprisingly, RE and correlation were highest for Z100 (although n is only 15). RE values based on historical data were clearly higher than the annually averaged RE_{median} time series from the SSV from the same time period. It must be noted that the validation data were from regions with a dense station network and therefore with a high reconstruction skill in the SSV. When taking into account the spatial distribution of the validation stations, comparable results are found for the historical data and the SSV. Figures 10a–f show scatterplots of reconstructed anomalies versus historical data. The plots show a good overall agreement, although the variability is underestimated (because of the least squares fitting). The comparison confirms the result from the SSV. The reconstruction is generally better for GPH than for temperature, for which the correlation drops off at 300 hPa.

Figures 10g–k show time series of monthly anomalies for selected periods and two validation sites. The error bars show the uncertainty in the observations of ±1°C for both levels and ±30 and ±20 gpm for the 500- and 700-hPa levels, respectively. For the reconstructions the 95% confidence interval was determined from the SSV. The overall agreement is excellent. Extremes are well

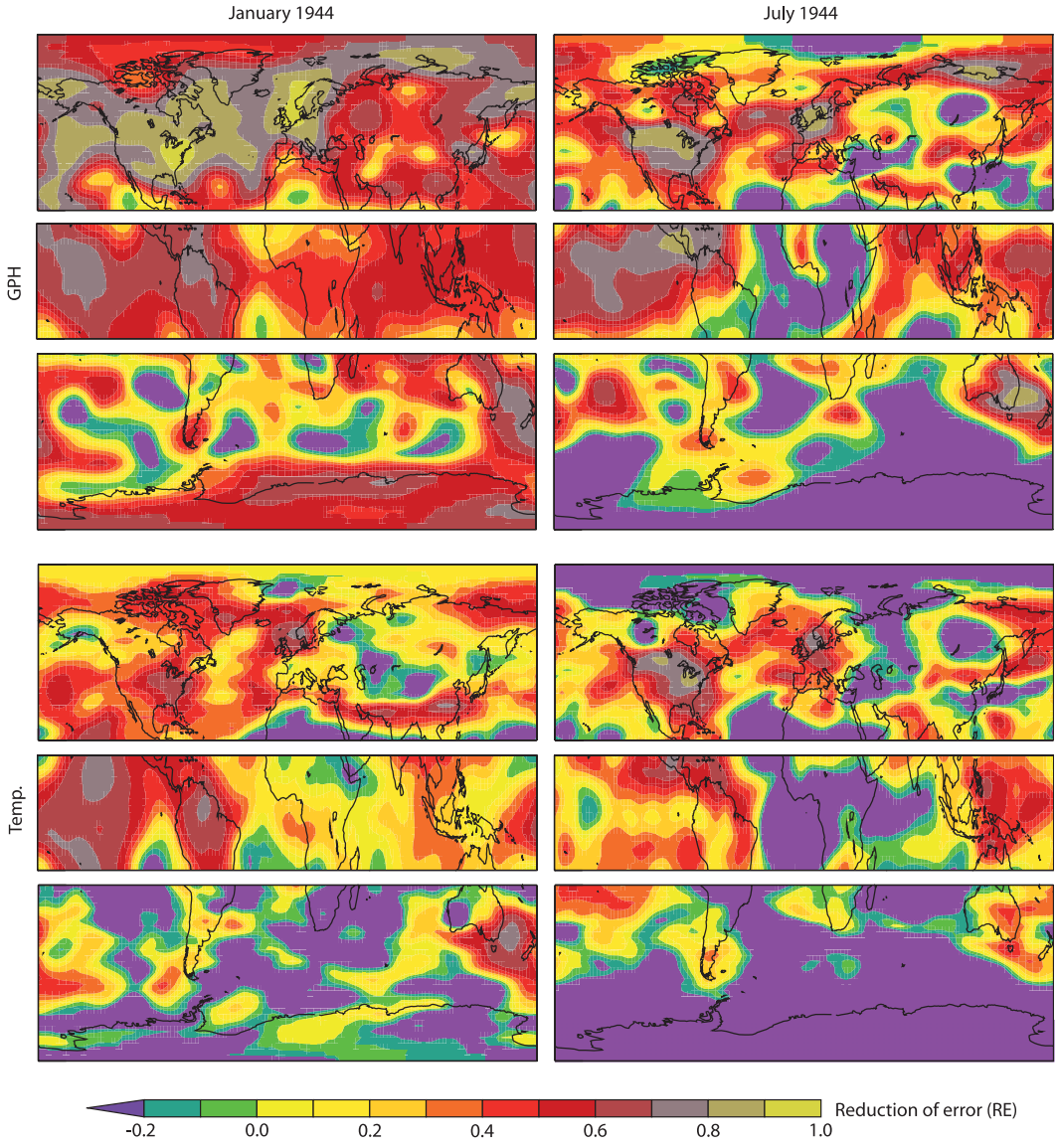

FIG. 8. As in Fig. 7, but for the months January and July 1944.

represented, although slightly underestimated because of our reconstruction approach. Data and reconstructions are mostly within each other's confidence intervals. For Oakland in 1942, a cold temperature and low-pressure bias appears. Because the number of predictors is large in these years, it is unlikely that such a bias is real. We therefore suspect that this is a remaining data problem.

Unfortunately, the data coverage in the Southern Hemisphere and parts of the tropics was very low. Therefore, no independent upper-air validation data with the required quality were available in these regions. However, it has already been shown in the SSV results that poor skill is to be expected for the Southern Hemisphere. Through the recovery of additional historical upper-air data, reconstructions and the corresponding validation will improve in the future.

The reconstruction agreed well with independent reconstructions from Schmutz et al. (2001) and Brönnimann and Luterbacher (2004). All levels for both reconstructions showed correlations above 0.8 for the pooled grid cells for each level. For the reconstruction by Schmutz et al. (2001), the correlation slightly dropped from 0.85 in the 700-hPa level to 0.81 in the 300-hPa level. Because Schmutz et al. (2001) do not include any upper-air data, this was expected. The reconstructions reproduce the 1940–42 anomalies at upper levels (related to El Niño) shown by Brönnimann et al. (2004).

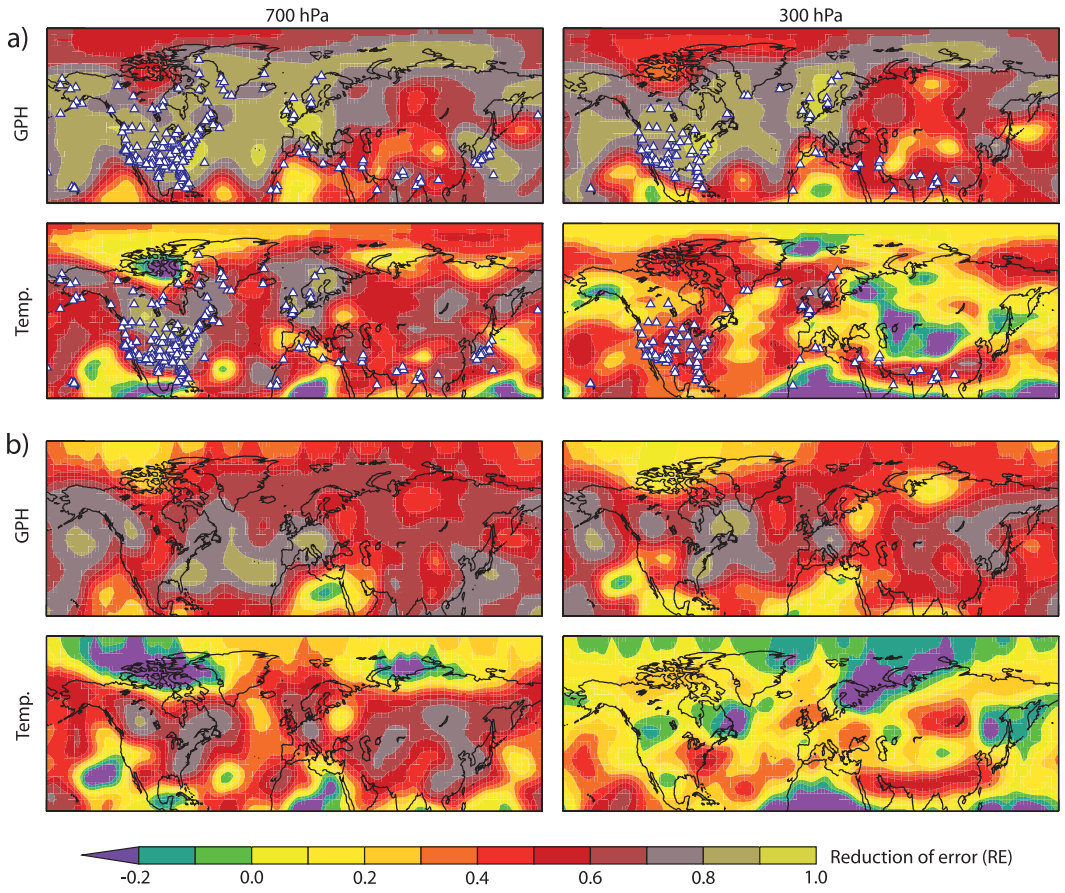

FIG. 9. RE fields for GPH and temperature for January 1944 in the NH, with (a) all available upper-air predictors included—white triangles represent (left) all available upper-air predictors for the specified month and (right) upper-air predictors available only above the 700-hPa level—and (b) only surface predictors.

Comparisons were also performed with the historical reanalysis of G. P. Compo et al. (2010, unpublished manuscript). Detailed results from these comparisons will be presented in another paper.

d. Validation with historical total ozone data

The results of the validation of the reconstructions against total ozone are shown in Fig. 11. To calculate the correlations, the reconstructions were merged with ERA-40 and the mean seasonal cycle from November 1978 to October 1994 (before the gap in the TOMS sensor was subtracted from all the series). Then, correlations were calculated separately for the pre-1957 and post-1958 period. Note that the correlations with total ozone show a characteristic profile: correlations increase in strength from the surface to the upper troposphere; they reach a maximum at about 200 hPa; and above that level, the correlation with temperature changes sign.

In general, the correlation profiles from the reconstructions (dashed lines) agreed well with those from the reanalysis period. Deviations can have three causes: 1) the low number of historical total ozone data (to account for this, the region of insignificant correlation for the historical period is shaded), 2) a low skill of the reconstructions, and 3) inaccurate total ozone data. Three of the total ozone series have been homogenized and can be considered to have a higher quality: Arosa, Oxford, and Tromsø. In fact, the agreement between the correlation profiles was excellent for these sites (note that lower correlations were expected for Tromsø; see solid lines). Good agreement was also found for New York. Results were particularly interesting for Canberra (1929–32) and Shanghai (1932–42), which represent periods and regions for which no upper-air data was included in the reconstructions (though the historical total ozone series are clearly of lower quality). Both sites showed significant correlations in the troposphere, confirming that the reconstructions have skill even in these cases.

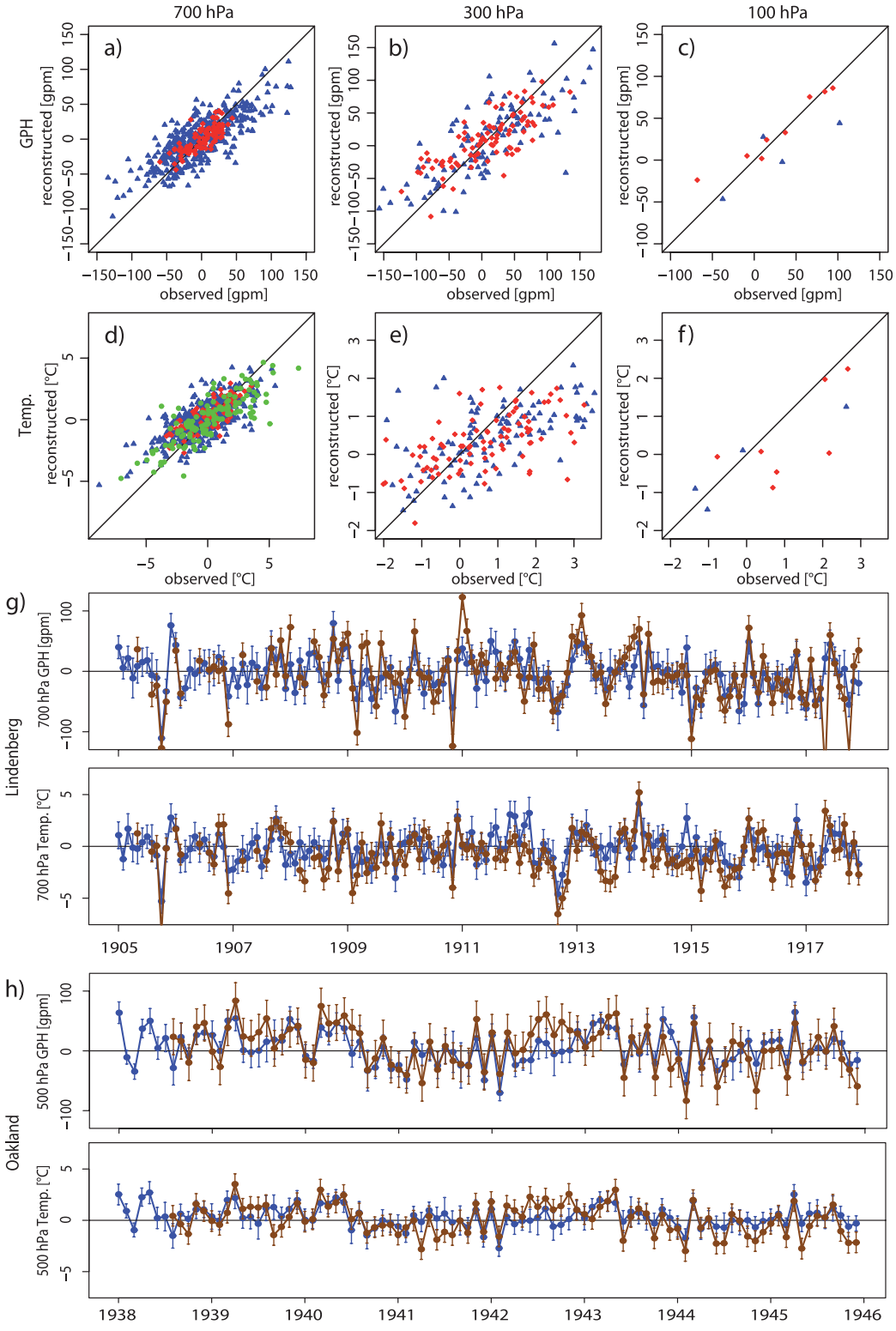

FIG. 10. Observed and reconstructed anomalies of (a)–(c) GPH and (d)–(f) temperature at (left to right) 700, 300, and 100 hPa for Ellendale (green), Lindenberg (blue), and Oakland (red). Time series of observed (brown) and reconstructed (blue) anomalies of (g) geopotential height and temperature at 700 hPa at Lindenberg for 1905–17 and (h) at 500 hPa at Oakland for the period 1938–45. Error bars give the assumed uncertainty of the observations and the 95% confidence intervals for the reconstructions. Anomalies are with respect to 1961–90.

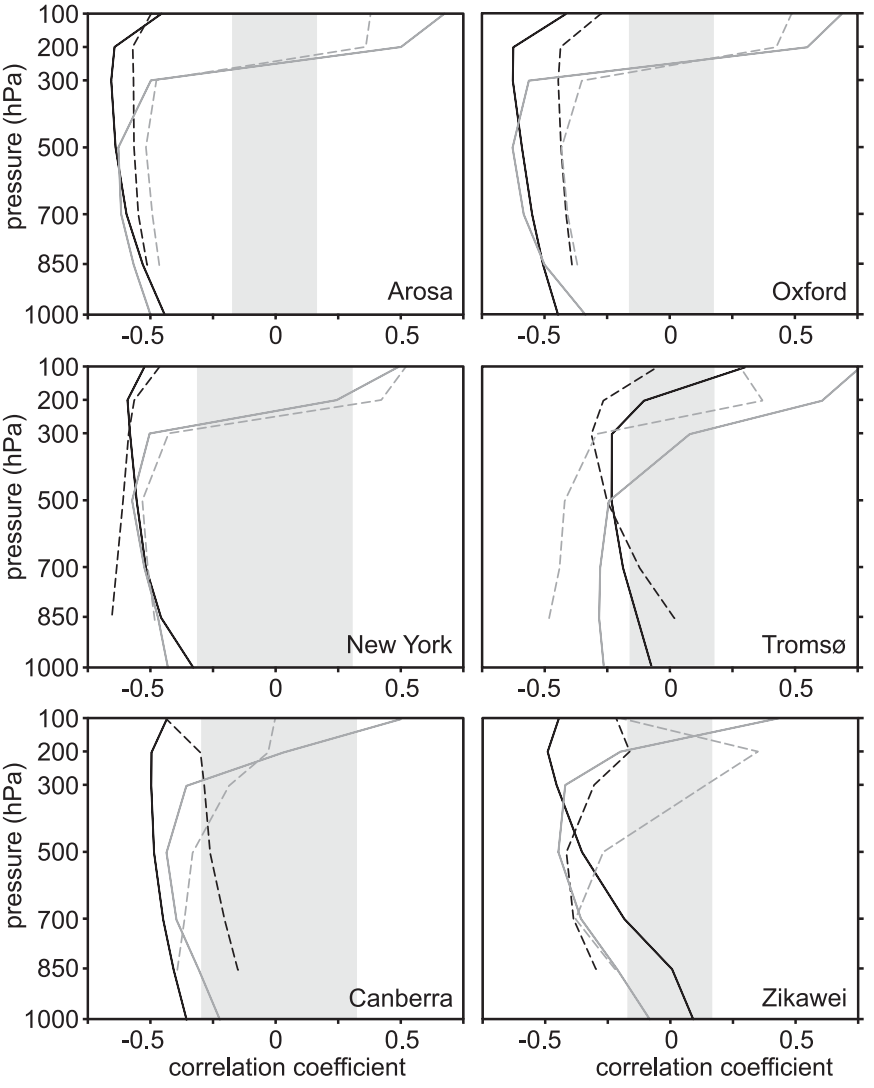

FIG. 11. Profiles of the correlation coefficient between total ozone and local upper-level variables for six sites. Gray lines are for temperature, black lines for GPH. Solid lines represent the period 1957–2001, dashed lines represent the reconstruction period. Areas were the correlation is insignificant in the reconstruction period are shaded.

5. An analysis of reconstructed fields

In this section, an analysis of selected fields is presented to point to possible applications of the reconstructions. Anomaly fields (with respect to the 1961–90 mean annual cycle) are shown for winter [December–February (DJF)] 1925/26, summer [June–August (JJA)] 1936, and winter (DJF) 1936/37. This selection was made because surface temperature and station temperature anomalies for different heights over the United States for the same months were already presented in a previous study by Ewen et al. (2008b, their Figs. 9–11), allowing a direct comparison. During the summer of 1936, in the middle of the decade of the "Dust Bowl" drought, high precipitation deficits and surface temperature anomalies occurred over the central United States. The winters 1925/26 and 1936/37 represent extreme positive and negative values of the Pacific–North America (PNA) index, respectively (see also Ewen et al. 2008a). Figure 12 shows the reconstructed fields of GPH and temperature for the levels 700, 300, and 100 hPa in the Northern Hemisphere. Shading denotes areas with poor skill (RE < 0.2) based on the SSV results.

On the 300-hPa level in both winters, a clear PNA-like pattern appears with the two north–south-oriented centers over the North Pacific and the two northwest–southeast-oriented centers over North America. In the winter of 1925/26, the latter were slightly rotated counterclockwise, whereas in 1936/37 the northern center over

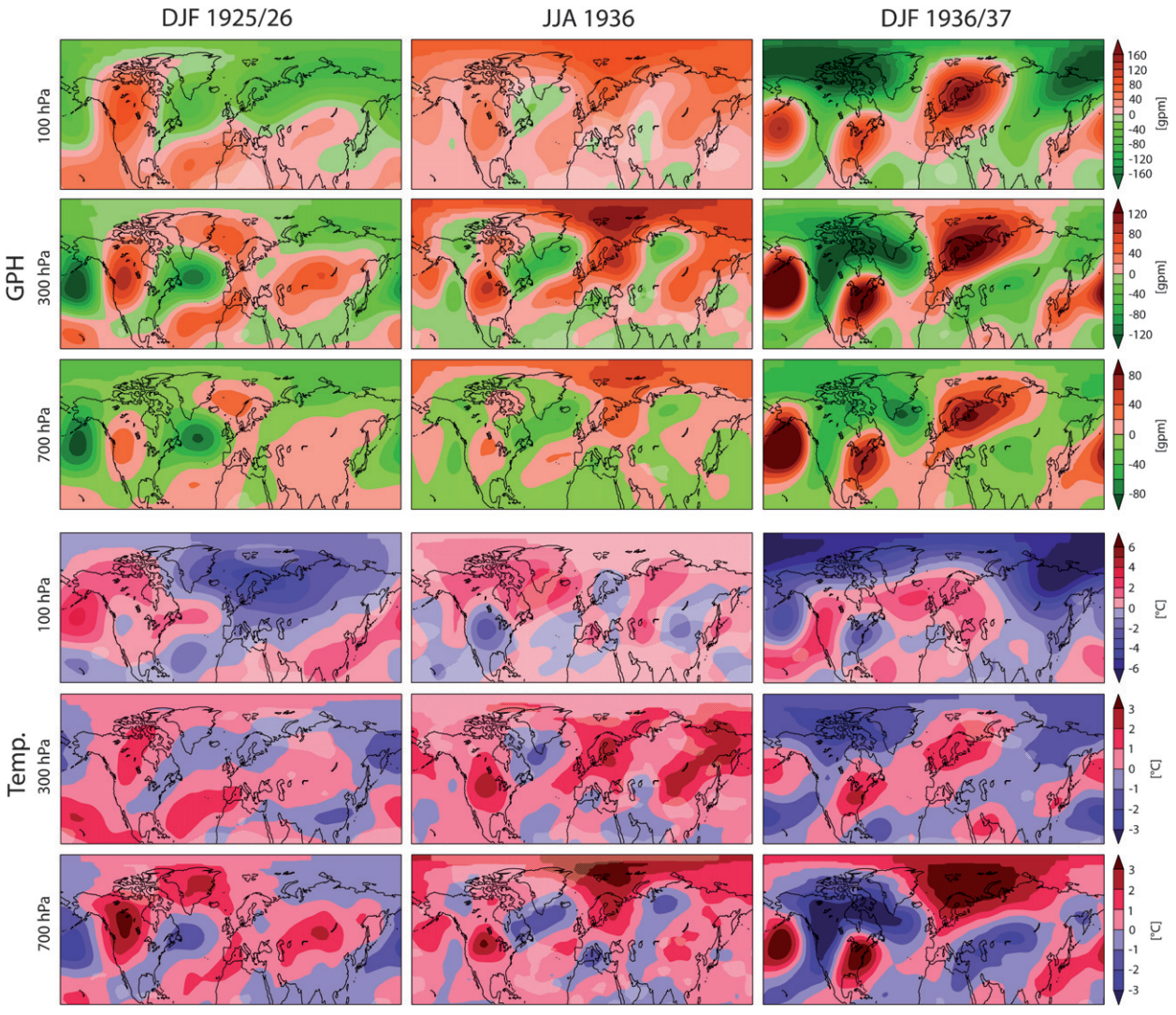

FIG. 12. Reconstructed anomaly fields of (bottom to top in each section) 700-, 300-, and 100-hPa (top three rows) GPH and (bottom three rows) temperature for (left) DJF 1925/26, (middle) JJA 1936, and (right) DJF 1936/37. Light shading denotes areas with RE < 0.2. Anomalies are with respect to the 1961–90 period.

North America extended into the Atlantic. Closer to the surface, the PNA signal becomes weaker but the pattern remains the same. Besides the strong PNA-like pattern over the United States, a prominent positive anomaly over Scandinavia is visible. This feature was potentially connected to increased transport of warm air into the Arctic region, seen in the lower troposphere (see, e.g., Grant et al. 2009a). On the 100-hPa level, a weak (1925/26) or strong (1936/37) negative anomaly over the pole is observable, pointing to a stronger than normal polar vortex. The GPH feature in the Arctic is consistent with the cold temperatures in the Arctic stratosphere in the same years (1925/26 and 1936/37). Although the reconstruction approach was designed to capture the large-scale (global) features, regional features such as the temperature gradient in the lower troposphere (see Ewen et al. 2008b) over the United States from warm in the southeast to cold in the northwest in the winter 1936/37 were well captured. Even the weakened gradient with height depicted by the single stations was present in the reconstructions. The same gradient in observations and reconstructions but with opposite sign was apparent in 1925/26.

In the summer of 1936, a positive GPH anomaly on the 300-hPa level over the United States was present. Midtropospheric ridging has already been linked to later Midwest droughts by Namias (1982) but was never shown to have actually occurred during the 1930s. Because this feature extended into the lower troposphere, a dynamical cause is likely. In the case of a mainly thermally driven feature, a negative anomaly would be

expected near the surface. The GPH anomaly is embedded into a wave train with further positive nodes over the Aleutian Islands, northern Europe, and eastern Asia, and negative nodes in the eastern North Pacific, northwestern North Atlantic, and Russia. The surface temperature anomaly over the United States in the summer of 1936 (see Ewen et al. 2008b) extended into the middle troposphere. At the same time, warmer than normal temperatures were observed over the Arctic region in the lower troposphere, possibly as a consequence of the positive GPH anomaly over northern Europe that was mentioned earlier.

6. Conclusions

Global monthly mean fields of temperature and GPH up to 100 hPa have been reconstructed for the period 1880–1957. For the reconstruction a statistical model was set up in the calibration period (1957–2002) linking the predictand and the predictor dataset. The derived model and coefficients were applied to historical predictor data in the reconstruction period. As predictand, the ERA-40 reanalysis was used. The predictor consisted of surface data such as gridded SLP and surface station temperature and upper-air measurements taken by radiosondes, kites, aircraft, and pilot balloons from the historical period before 1958. A total of 15 394 upper-air predictor variables were used. To factor in the lowered data quality in the reconstruction period, the predictor data in the calibration period was perturbed with normally distributed noise.

The quality of the reconstruction was checked based on a statistical split-sample approach, independent historical upper-air measurements, historical total ozone data, and independent reconstructions. Validation experiments revealed a good overall quality. However, there was a large spread in the skill of the reconstructions. When working with the reconstructions, the seasonal and spatial variability of the skill has to be considered.

The best reconstructions were found for GPH and the Northern Hemisphere in the winter season, and lower levels were better reproduced than higher levels. The results for the tropics and the Southern Hemisphere should be interpreted with care. Over the ocean the model often showed poor skill. It is important that the reconstructed fields are used only after due consideration of the corresponding RE fields.

The application of the reconstructed fields to selected months showed that the reconstructions are suitable for studying the large-scale circulation. Through the recovery of historical upper-air data, the skill of the reconstructions could be improved in the future, especially in the tropics and in the Southern Hemisphere.

Acknowledgments. This work was supported by the Swiss National Science Foundation, Project ''past climate variability from an upper-level perspective.'' We wish to thank all data providers, especially Roy Jenne (NCAR) and Tom Ross (NOAA/NCDC) as well as Météo-France for providing pilot balloon data. Wolfgang Adam (DWD, German Weather Service) provided the historical data from Lindenberg that was used for the validation.

REFERENCES

Allan, R., and T. Ansell, 2006: A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004. *J. Climate,* **19,** 5816–5842.

Ansell, T. J., and Coauthors, 2006: Daily mean sea level pressure reconstructions for the European–North Atlantic region for the period 1850–2003. *J. Climate,* **19,** 2717–2742.

Bengtsson, L., S. Hagemann, and K. I. Hodges, 2004: Can climate trends be calculated from reanalysis data? *J. Geophys. Res.,* **109,** D11111, doi:10.1029/2004JD004536.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperatures changes: A new dataset from 1850. *J. Geophys. Res.,* **111,** D12106, doi:10.1029/2005JD006548.

Brönnimann, S., 2003a: Description of the 1939–1944 upper-air dataset (UA39-44) version 1.1. University of Arizona, 40 pp.

——, 2003b: A historical upper-air dataset for the 1939–1944 period. *Int. J. Climatol.,* **23,** 769–791.

——, and J. Luterbacher, 2004: Reconstructing Northern Hemisphere upper-level fields during World War II. *Climate Dyn.,* **22,** 499–510.

——, and J. Staehelin, 2004: Total ozone prior to the International Geophysical Year (IGY). *Proc. XX Quadrennial Ozone Symp.,* Vol. 1, Kos, Greece, International Ozone Commission, 304–305.

——, J. C. Cain, J. Staehelin, and S. F. G. Farmer, 2003: Total ozone observations prior to the IGY. II: Data and quality. *Quart. J. Roy. Meteor. Soc.,* **129,** 2819–2843.

——, J. Luterbacher, J. Staehelin, T. M. Svendby, G. Hansen, and T. Svenoe, 2004: Extreme climate of the global troposphere and stratosphere in 1940–42 related to El Niño. *Nature,* **431,** 971–974.

Casty, C., H. Wanner, J. Luterbacher, J. Esper, and R. Böhm, 2005: Temperature and precipitation variability in the European Alps since 1500. *Int. J. Climatol.,* **25,** 1855–1880.

Collins, W., and Coauthors, 2006: The Community Climate System Model version 3 (CCSM3). *J. Climate,* **19,** 2122–2143.

Cook, E. R., K. R. Briffa, and P. D. Jones, 1994: Spatial regression methods in dendroclimatology—A review and comparison of two techniques. *Int. J. Climatol.,* **14,** 379–401.

Durre, I., R. S. Vose, and D. B. Wuertz, 2006: Overview of the integrated global radiosonde archive. *J. Climate,* **19,** 53–68.

Eskridge, R., A. Alduchov, I. Chernykh, Z. Panmao, A. Polansky, and S. Doty, 1995: A Comprehensive Aerological Reference Data Set (CARDS): Rough and systematic errors. *Bull. Amer. Meteor. Soc.,* **76,** 1759–1775.

Ewen, T., S. Brönnimann, and J. Annis, 2008a: An Extended Pacific–North American index from upper-air historical data back to 1922. *J. Climate,* **21,** 1295–1308.

——, A. Grant, and S. Brönnimann, 2008b: A monthly upper-air dataset for North America back to 1922 from the *Monthly Weather Review*. *Mon. Wea. Rev.,* **136,** 1792–1805.

Gong, D. Y., H. Drange, and Y. Q. Gao, 2006: Reconstruction of Northern Hemisphere 500-hPa geopotential heights back to the late 19th century. *Theor. Appl. Climatol.,* **90,** 83–102.

Grant, A. N., S. Brönnimann, T. Ewen, T. Griesser, and A. Stickler, 2009a: The early twentieth century warm period in the European Arctic. *Meteor. Z.,* **18,** 425–432.

——, ——, ——, and A. Nagurny, 2009b: A new look at radiosonde data prior to 1958. *J. Climate,* **22,** 3232–3247.

Hansen, G., and T. Svenøe, 2005: Multilinear regression analysis of the 65-year Tromsø total ozone series. *J. Geophys. Res.,* **110,** D10103, doi:10.1029/2004JD005387.

Hansen, J., R. Ruedy, J. Glascoe, and M. Sato, 1999: GISS analysis of surface temperature change. *J. Geophys. Res.,* **104,** 30 997–31 022.

Jones, P. D., T. M. L. Wigley, and K. R. Briffa, 1987: Monthly mean pressure reconstructions for Europe (back to 1780) and North America (to 1858). U.S. Dept. of Energy Carbon Dioxide Research Division, Tech. Rep. TRO37, 99 pp.

Kington, J. A., 1975: The construction of 500-millibar charts for the eastern North Atlantic–European sector from 1781. *Meteor. Mag.,* **104,** 336–340.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.,* **82,** 247–267.

Klein, W. H., and Y. Dai, 1998: Reconstruction of monthly mean 700-mb heights from surface data by reverse specification. *J. Climate,* **11,** 2136–2146.

Lanzante, J. R., S. A. Klein, and D. J. Seidel, 2003: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. *J. Climate,* **16,** 224–240.

Luterbacher, J., and Coauthors, 2002: Reconstruction of sea level pressure fields over the eastern North Atlantic and Europe back to 1500. *Climate Dyn.,* **18,** 545–561.

——, D. Dietrich, E. Xoplaki, M. Grosjean, and H. Wanner, 2004: European seasonal and annual temperature variability, trends, and extremes since 1500. *Science,* **303,** 1499–1503.

Mitchell, T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.,* **25,** 693–712.

Namias, J., 1982: Anatomy of Great Plains protracted heat waves (especially the 1980 U.S. summer drought). *Mon. Wea. Rev.,* **110,** 824–838.

Parker, D. E., M. Gordon, D. P. N. Cullum, D. M. H. Sexton, C. K. Folland, and N. Rayner, 1997: A new global gridded radiosonde temperature database and recent temperature trends. *Geophys. Res. Lett.,* **24,** 1499–1502.

Pauling, A., J. Luterbacher, C. Casty, and H. Wanner, 2005: Five hundred years of gridded high-resolution precipitation reconstructions over Europe and the connection to large-scale circulation. *Climate Dyn.,* **26,** 387–405.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.,* **108,** 4407, doi:10.1029/2002JD002670.

Santer, B. D., and Coauthors, 2004: Identification of anthropogenic climate change using a second-generation reanalysis. *J. Geophys. Res.,* **109,** D21104, doi:10.1029/2004JD005075.

Schmutz, C., D. Gyalistras, J. Luterbacher, and H. Wanner, 2001: Reconstruction of monthly 700-, 500-, and 300-hPa geopotential height fields in the European and eastern North Atlantic region for the period 1901–1947. *Climate Res.,* **18,** 181–193.

Simmons, A. J., and Coauthors, 2004: Comparison of trends and low-frequency variability in CRU, ERA-40, and NCEP–NCAR analyses of surface air temperature. *J. Geophys. Res.,* **109,** D24115, doi:10.1029/2004JD005306.

Smith, T. M., and R. W. Reynolds, 2004: Improved extended reconstruction of SST (1854–1997). *J. Climate,* **17,** 2466–2477.

Staehelin, J., and Coauthors, 1998: Total ozone series at Arosa (Switzerland): Homogenization and data comparison. *J. Geophys. Res.,* **103,** 5827–5841.

Stickler, A., and Coauthors, 2010: The Comprehensive Historical Upper-Air Network (CHUAN). *Bull. Amer. Meteor. Soc.,* in press.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Vogler, C., S. Brönnimann, J. Staehelin, and R. E. M. Griffin, 2007: The Dobson total ozone series of Oxford: Re-evaluation and applications. *J. Geophys. Res.,* **112,** D20116, doi:10.1029/2007JD008894.

Xoplaki, E., J. Luterbacher, H. Paeth, D. Dietrich, N. Steiner, M. Grosjean, and H. Wanner, 2005: European spring and autumn temperature variability and change of extremes over the last half millennium. *Geophys. Res. Lett.,* **32,** L15713, doi:10.1029/2005GL023424.